

Automatic Identification of Patterns for ProDom families using Pratt

Florence Servant¹, Daniel Kahn¹, Jérôme Gouzy¹, Florence Corpet², Stig Dommarsnes³, Inge Jonassen³

¹ INRA/CNRS LBM RPM, BP 27, 31326 Castanet-Tolosan, France. ² INRA LGC, BP 27, 31326 Castanet-Tolosan, France. ³ Dept of Informatics, Univ. Of Bergen, Norway

The InterPro project (1) aims the creation of a new database of protein families, domains and functional sites. Each InterPro entry merges information from several other databases: PROSITE, PRINTS, Pfam and ProDom (2). The first ones are based on biological expertise whereas ProDom is built automatically from SWISS-PROT and TrEMBL sequences. One of the aims of InterPro is to expand the number of characterised families and to develop better signatures (patterns) for the families. We have selected novel protein families from the ProDom database and systematically searched for patterns applying the Pratt sequence pattern discovery tool (3,4) to each of them, in order to prepare their integration into InterPro.

The novel families not referenced in PROSITE 15.0 have been selected from ProDom on the following basis:

- they must have at least 2 members;
- they must contain at least one single-domain protein from SWISS-PROT (the domain defined by ProDom has not been artificially truncated, as it comes from a well characterised natural sequence);
- the single-domain sequence must be shorter than 500 amino acids (larger modules are probably multi-domain);
- the similarity between the most distant sequences lies between 10% identity (family homogeneity), and 90% identity (remove highly redundant families).

With these criteria, 1893 families were selected from the ProDom 99.1 database as novel families.

A method for large scale pattern searching has been developed in order to find patterns for each of the selected families. It builds on the Pratt program that allows automatic discovery of patterns matching at least a minimum number of a given set of unaligned sequences. The search performed by Pratt is governed by user-defined parameters, most importantly; the minimum number of sequences that a pattern should match, and the constraints on the patterns to be considered. The patterns produced are regular expressions of the same form as those used in the PROSITE database. The output from Pratt is a list of patterns satisfying the constraints, reported with their information contents (IC) and number of matching sequences. The higher the IC of a pattern, the less likely it is that a sequence will match it « by chance ». For the purpose of this project patterns are required to have a minimum IC in order to obtain family specificity. We have developed a « wrapper » for Pratt WPratt that runs Pratt repeatedly on each sequence set (family) adjusting parameters as necessary in order to obtain a pattern that has an IC high enough to be specific for the family and at the same time matching as many family sequences as possible. WPratt also scans the patterns with sufficiently high IC against the complete ProDom database to evaluate their specificity.

Another solution should have been to use an algorithm which search for patterns on a set of pre-aligned sequences, as implemented in the EMOTIF program. However, this method depends on the quality of the multiple alignment, that could be tricky for big families, or when sequences in the family are heterogeneous in length. Thus, we have preferred to use Pratt, without taking account the multiple alignments given by ProDom.

The patterns identified will be integrated into the InterPro database in order to help the semi-automatic annotation of SWISS-PROT. Thus, they must be selected on strict criteria: they must be discriminant enough, but at the same time patterns derived from too dense sequence sets will not generalise to a wider family. For this reason, patterns have been selected based on their IC by only including patterns with IC between 15 and 70 which is the range into which all PROSITE patterns fall.

For each pattern, the number FN of false negatives (family members that do not match the pattern), as well as the number FP of false positives (other sequences that match the pattern) is determined. The most promising patterns have been selected as followed (n is the family size): if $FN/n > 0.1$, keep patterns with $FP=0$. Otherwise, accept patterns with $FP < n/10$. Among the 1893 families selected from ProDom 99.1, 347 have at least 10 members ($n > 10$). With our pattern selection criteria, 239 families out of 347 have at least one promising pattern computed by WPratt. These informative and selective patterns will be soon integrated into InterPro.

References

1 InterPro Web Site: <http://www.ebi.ac.uk/interpro/>

2 CORPET F., SERVANT F., GOUZY J. & KAHN D. (2000): ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research*, **28**, 267-269.

3 JONASSEN I. (1997): Efficient discovery of conserved patterns using a pattern graph. *CABIOS*, **13**, 509-522

4 JONASSEN I., COLLINS J.F. & HIGGINS D.G. (1995): Finding flexible patterns in unaligned protein sequences. *Protein Sci.*, **4**, 1587-1595.